Y/CSD/INF-98/6

Y-12

OAK RIDGE
Y-12
PLANT

Report of Official Foreign Travel to France May 8–27, 1998

James David Mason
Internet, SGML, and Integration Services
Information Technology Services

June 11, 1998

Prepared by the
Oak Ridge Y-12 Plant
Oak Ridge, Tennessee 37831
managed by
Lockheed Martin Energy Systems, Inc.
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-84OR21400

MANAGED BY
LOCKHEED MARTIN ENERGY SYSTEMS
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

Y/CSD/INF-98/6

Report of Official Foreign Travel to France
May 8–27, 1998

James David Mason
Internet, SGML, and Integration Services
Information Technology Services

June 11, 1998

Prepared by the
Oak Ridge Y-12 Plant
Oak Ridge, Tennessee 37831
managed by
Lockheed Martin Energy Systems, Inc.
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-84OR21400

Report of Official Foreign Travel to France
May 8–27, 1998

James David Mason

Abstract

The Department of Energy (DOE) has moved ever more rapidly towards electronic production, management, and dissemination of scientific and technical information. The World-Wide Web (WWW) has become a primary means of information dissemination. Electronic commerce (EC) is becoming the preferred means of procurement. DOE, like other government agencies, depends on and encourages the use of international standards in data communications.

Among the most widely adopted standards is the Standard Generalized Markup Language (SGML, ISO 8879:1986, FIPS 152), which DOE has selected as the basis of its electronic management of documents. Besides the official commitment, which has resulted in several specialized projects, DOE makes heavy use of coding derived from SGML, and its use is likely to increase in the future. Most documents on the WWW are coded in HTML (Hypertext Markup Language), which is an application of SGML. The World-Wide Web Consortium (W3C), with the backing of major software houses like Microsoft, Adobe, and Netscape, is promoting XML (eXtensible Markup Language), a class of SGML applications, for the future of the WWW and the basis for EC.

In support of DOE's use of these standards, I have served since 1985 as Convenor of the international committee responsible for SGML and related standards, ISO/IEC JTC1/WG4 (WG4). During this trip I convened the spring 1998 meeting of WG4 in Paris, France. I also attended a major conference on the use of SGML and XML. At the close of the conference, I chaired a workshop of standards developers looking at ways of improving online searching of electronic documents. **Note:** Since the end of the meetings in France, JTC1 has raised the level of WG4 to a full Subcommittee; its designator is now ISO/IEC JTC1/SC34.

WG4 maintains and continues to enhance several standards. In addition to SGML, which is the basis of HTML and XML, WG4 also works on the Document Style Semantics and Specification Language (DSSSL), which is the basis for the W3C's XSL (eXtensible Style Language, to be used with XML) and the Hypermedia/Time-based Document Structuring Language (HyTime), which is a major influence on the W3C's XLink (XML Linking Language). WG4 is also involved in work with the ISO's TC184, Industrial Data, on the linking of STEP (the standard for the interchange of product model data) with SGML.

In addition to the widespread use of the WWW among DOE's plants and facilities in Oak Ridge and among DOE sites across the nation, there are several SGML-based projects at the Y-12 Plant. My project team in Information Technology Services has developed an SGML-based publications system that has been used for several major reports at the Y-12 Plant and Oak Ridge National Laboratory (ORNL). SGML is a component of the Weapons Records Archiving and Preservation (WRAP) project at Y-12 and is the format for catalog metadata chosen for weapons records by the Nuclear Weapons Information Group (NWIG).

Supporting standards development allows DOE and Y-12 both input into the process and the opportunity to benefit from contact with some of the leading experts in the subject matter. Oak Ridge has been for some years the location to which other DOE sites turn for expertise in SGML and related topics.

Note: This report is in many ways a sequel to my most recent foreign trip report, ORNL/FTR-5800, which reported on the Spring 1996 meeting of ISO/IEC JTC1/SC18/WG8, the predecessor of WG4, in Munich, Germany. There have been also other meetings of WG4 during 1996 and 1997 that did not result in a foreign trip reports; copies of documentation for these meetings are available from the WG4 site on the WWW (http://www.ornl.gov/sgml/sc34/).

This report is available on the WG4 Web site at http://www.ornl.gov/sgml/sc34/document/1997.htm. Hyperlinks in the online report connect it to the documents it references on both the WG4 site and at other locations, particularly the W3C.

Introduction

In the Joint Technical Committee on Information Technology (JTC1) of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), the responsibility for standards in the area of Document Description and Processing Languages lies with ISO/IEC JTC1/WG4 (WG4), which I convene. **Note:** Since the end of the meetings in France, JTC1 has raised the level of WG4 to a full Subcommittee; its designator is now ISO/IEC JTC1/SC34.

One of WG4's standards—SGML—is among the most widely used of all ISO standards. The European Community adopted SGML for its official publications even before the standard was completed. It was made a Federal Information Processing Standard (FIPS 152) and was adopted many years ago by the U.S. Department of Defense (DoD) as part of its CALS (variously "Computer-aided Acquisition and Logistic Support" or "Commerce at Lightspeed" over the years) initiative. SGML was also adopted in 1990 by DOE, not simply because of DOE's close ties to DoD but even more because the standard can be an excellent basis for a system of capturing and managing the scientific and technical information developed by DOE's research facilities. The Nuclear Weapons Information Group (NWIG) has adopted SGML as the form for metadata in catalogs of weapons data at DOE sites.

SGML is the basis on which HTML (http://www.w3.org/TR/REC-html40/), the coding convention for most documents on the WWW, was built. Lately the W3C has been promoting a more flexible approach to coding systems they call XML (http://www.w3.org/XML/Activity), which is a potentially very large class of SGML applications that they believe will replace HTML for many documents on the WWW. Because HTML, as a single SGML application, has only one set of tags to identify information elements, developers of WWW content have been frustrated with its limitations. XML, which allows users to develop new SGML applications with elements and tags designed to reflect their particular information needs, is gaining wide acceptance. Both Microsoft and Netscape are supporting XML in their WWW browsers, and Microsoft, Adobe, and other major software houses plan to support it across their product lines.

The other projects under development in WG4 are also of interest to the government because they can be used with SGML to develop a comprehensive and powerful publications and information management systems. Among the most notable of these are DSSSL (ISO/IEC 10179:1996) and HyTime)ISO/IEC 10744:1992, 1997). The W3C is using these standards as the basis for a suite of simplified standards to support XML, just as the full standards support complex SGML applications. Because these standards have the backing of the major software houses, they are likely to have wide influence in DOE.

Spring Meeting of ISO/IEC JTC1/WG4, Paris, France

The meeting of WG4 was held at AIS, S.A., a major European vendor of SGML software and services. The attendance at the spring meeting of WG4 included 22 experts representing seven national bodies (Canada, France, Germany, Japan, Norway, the United Kingdom, and the United States), three external liaison bodies (SGML Users' Group, CERN, and ISO TC184/SC4, Industrial Data).

The opening plenary was held on Monday, 11 May 1998, with reports from national bodies, liaison organizations, and project editors. After status reports on the projects, the working group stayed mostly in a single session because of common interest in revision of SGML. There were breakout sessions for HyTime and Topic Navigation Maps (A HyTime-based standard) during the week.

At this meeting WG4 received a liaison delegation from ISO TC184/SC4, which is responsible for STEP, the standard for industrial product model data representation and exchange. WG4 had already designated a liaison to this group at a previous meeting. One of the oldest and largest areas of SGML application has been industrial documentation (e.g., product specifications, operating and maintenance manuals). TC184/SC4 (http://www.nist.gov/sc4/) is responsible for the STEP standard and EXPRESS, the language that supports it. Because the same product will often have data concerning it in both SGML and STEP/EXPRESS, the group has begun a new project to investigate interconnections between the two systems of data representation.

The W3C is also developing a program for liaisons to international standards bodies. A longtime member of WG4, Mr. Rudolf Riess of DEC (Compaq Computer), has been assigned to the W3C to support this program, and he also attended the meeting.

Project Meetings

SGML

WG4's oldest ISO standard, SGML (ISO 8879:1986), is evolving with changes in information technology. The process began with the completion of DSSSL and the republication of HyTime. Since then, the W3C has chosen a subset of those three standards as the basis for XML, XSL, and XLink, which are attracting wide attention from both users and software vendors. WG4 has published two Technical Corrigenda (TCs) to SGML to support internationalization of text (through UNICODE/ISO 10646) and to formalize some of the constraints imposed on applications by XML.

At this meeting, WG4 completed work on a third TC to extend the support for XML and looked at methods for modularizing the definitions of document structures (including proposals similar to the XML "Namespace" proposal). The W3C's XML groups have been investigating alternative notations for what SGML calls Document Type Definitions (DTDs); the TC will support not only the W3C's notations but also those that may emerge from the STEP/EXPRESS world.

WG4 is also working on an ISO version of HTML (in cooperation with the W3C). WG4 reviewed the latest draft from the editors and made comments in preparation for sending it out for a final phase of balloting.

HyTime and Related Topics

HyTime (ISO/IEC 10744), a standard for manipulating multimedia data and hyperlinking diverse forms of information, continues to gain attention. WG4 is completing a second TC and an amendment to the standard. Topic Navigation Maps (ISO/IEC CD 13250) is a specific application of HyTime that emphasizes the ability to add navigational tools to existing bodies of information. Much of the work in

Paris centered around this project. The concept of "SGML Property Sets," which has evolved from DSSSL and HyTime, may also become the basis of the linking to STEP/EXPRESS data representations.

Results of the Meeting

WG4 is pleased that its standards continue to attract attention and new applications. The group is particularly pleased by the rapid growth of XML and will concentrate its efforts on supporting that class of applications.

The *Recommendations* of the WG4 Meeting (http://www.ornl.gov/sgml/wg8/document/1972rec.htm) and the WG4 *Programme of Work* (http://www.ornl.gov/sgml/wg8/document/1971bpro.htm) are available on line as formal statements of the accomplishments of the meeting. The WG4 library also includes the *Annual Report* (http://www.ornl.gov/sgml/wg8/document/1970ann.htm) that will be presented to the JTC1 Plenary in Sendai, Japan.

ISO/IEC JTC1 Plenary

The annual plenary meeting of WG4's parent committee, ISO/IEC JTC1, is being held in Sendai, Japan, the week of 2 June 1998. Although I, as the WG4 Convenor, am expected to attend the JTC1 Plenary, this year constraints in time and funding caused me to return after the WG4 meeting and submit my report through Mr. Rudolf Riess. Mr. Riess has reported from Japan that JTC1 has decided to raise WG4's status from working group to subcommittee, so that in the future it will be known as JTC1/SC34.

Conference: *SGML/XML Europe '98*

The Graphic Communications Association (GCA, an affiliate of Printing Industries of America) has been a supporter of SGML and its applications from the earliest days. Their conferences on SGML-related topics had grown steadily over the years, but the arrival of first HTML and then XML has caused an explosion of participation in both North America and Europe. (My first European conference, in 1984, was held in one large lecture hall at Oxford University; this year's conference put a stress on one of the largest hotels in Paris.) Some sense of how the world of information is moving can be gained from the keynote speakers: representatives of Microsoft, Adobe, Sun, and the Organization for Economic Cooperation and Development (OECD) showed how the major software and hardware suppliers are rushing to adopt XML, and the user community can hardly wait for them to release the products. The title of a later Microsoft presentation gave a hint of where that company is headed: "XML Everywhere." (Internet Explorer 4.0 includes an XML processor. Microsoft is giving away source code in Java to another processor, and they demonstrated the XML functions in Office 8. It looks, for example, as though XML+XSL will replace RTF as the base storage format for Word, and other components will also use XML in some way.)

The conference, which generally had six concurrent tracks, was too vast for me to absorb by myself (I have the proceedings in both paper and electronic form for anyone wanting to inspect them).

The track on STEP/SGML harmonization drew steady attendance throughout the conference: Many organizations seem to be interested in unifying their product models, combining manufacturing data with documentation and records keeping. The issues raised in this track also came up in others: managing documentation life cycles, managing document components in the same way a product-data manager handles manufacturing data, managing information content. Quite a few vendors were promoting component-management systems. Data modelling has long been an interest of SGML users, and several new approaches, such as the XML-Data proposal in the W3C (http://www.w3.org/TR/1998/NOTE-XML-data/), were attracting attention. Modelling is key to effective metadata use, and it is also central to linking

engineering and manufacturing processes to documentation. Although Electronic Data Interchange (EDI) has been around for many years as a component of EC, it has depended largely on proprietary protocols (sometimes loosely packaged as "standards"). The SGML community has long asserted that EDI transaction sets were perfectly suited for SGML representation. Now, because of the widespread interest in XML, the early promise of SGML in this area seems to be coming to pass (http://www.xmledi.net/). The processing of healthcare data, which covers a wide range of subjects from patient records (and related privacy issues) to aspects of EC, continues to attract attention. The Europeans may actually be implementing SGML/XML solutions more rapidly than U.S. organizations.

The conference was quite lively, and interest in the SGML/XML world is growing explosively.

ISO/IEC JTC1/WG4 Workshop: Metadata and Online Searching

Observing the proliferation of online information resources and the limitations of current search techniques, WG4 proposed a workshop (http://www.ornl.gov/sgml/wg4/document/1946.htm) on "Guidelines for accessing data and metadata represented in SGML from databases, knowledge bases and search tools." We hoped that the workshop would bring together standards developers (and implementers) from a wide variety of communities and begin to establish a common framework for both coding and semantics to improve information delivery. In proposing the workshop, WG4 recognized that brute-force search techniques (used by typical WWW search engines like Yahoo and Altavista) cannot refine data sufficiently, and keyword searching is both too poorly supported and too inconsistent in application to be of much help. Our hope is to supplement these techniques with a common framework of metadata (data about data—a library card catalog, for example, contains metadata about the books on the shelves) and means of representing and exchanging it. A commonly recognized starting point for metadata discussions is the "Dublin Core" that has evolved out of discussions in the library and information-science community (http://purl.oclc.org/metadata/dublin_core/). The W3C is working on the Resource Description Framework (RDF, http://www.w3.org/Metadata/) as an XML metadata application, starting with the Dublin Core. (In the DOE community, the NWIG metadata set bears considerable similarity to Dublin Core.)

The workshop was hosted by the GCA at the conclusion of their conference *SGML/XML Europe '98*. Participation in the workshop was very good, with over 40 representatives from ISO committees, the W3C, and major software houses. Among the ISO groups represented were JTC1/SC32 (responsible for SQL), JTC1/SC24 (computer graphics), TC46 (Information and Documentation) and TC184/SC4, in addition to most of the projects in WG4 (including most of the editors of SGML, HyTime, DSSSL, and Topic Navigation Maps). The W3C was represented by members of the editorial teams for XML, XSL, RDF, and XML-Data. Most of the day was occupied by presentations from the various bodies about both their requirements for metadata and the approaches they have considered. The day ended with a discussion of technical considerations deserving further study.

There seemed to be general consensus that it might be possible to accept a common set of metadata, such as the Dublin Core, but that an extension capability was also clearly necessary to support searching of individualized databases. There is general agreement that SGML/XML tools should be central to the expression of metadata, even in areas like computer graphics that have traditionally used other means of encoding data (the convergence of STEP and SGML has already been noted; the developers of CGM [Computer Graphics Metafile] are also interested in such issues). At the same time, there is a recognition that traditional databases, with SQL-based query systems, have reached a high level of development. Consequently, it is desirable to have a flexible conversion between SGML/XML metadata and the tuples (multiple pieces of information seen as a unit) understood by traditional database management systems.

The W3C community brings with it additional issues related to online operations, such as the need to support “trusted metadata” (i.e., metadata with digital signatures) and the goal of supporting intelligent agents for metadata searching and manipulation.

The details of expression and extensibility are still quite open to discussion. The developers of RDF are particularly concerned with conversion of metadata from XML to tuples (the basic construct of RDF can be expressed as a 3-tuple or “triple”). Extensibility is gained by combining metadata from various sources using the proposed XML “Namespace” mechanism (http://www.w3.org/TR/WD-xml-names). The notation for RDF tends towards hierarchical but otherwise straightforward XML.

Another approach, which seems at first glance to be quite different from that of RDF, comes from the HyTime community. One of the goals of HyTime was, from the beginning, to apply linking and navigation structures over data collections without necessarily intervening in their contents (this is the basic technique of Topic Navigation Maps). One of the HyTime editors demonstrated at the conference (and again briefly at the workshop) the ability to link into the contents of CGM graphics, even without the SGML additions that are being considered. A major HyTime concept is that of “architectural forms” that allow adding layers of information structure to an SGML (or XML) DTD without causing the users of the SGML application to have to change how they tag their documents.

Traditional SGML and XML thinking has been largely about the syntactic structures (e.g., tag sets and their usage rules) resulting from a DTD. Applications may invest tags with semantics (an HTML <A> tag causes hyperlinking on the WWW), or users may make semantic associations (a <P> tag marks a paragraph). For metadata to be meaningfully searchable, it must convey semantic, rather than syntactic, structures. The RDF approach is to approach semantic combination as an implied consequence of combining components drawn from varied namespaces. The cost is potentially verbose metadata structures. The HyTime proponents contend that their mechanisms allow the construction of “semantic groves” for searching with less impact on the creators and users of documents than the current RDF proposals. In exchange for the additional up-front labor of setting up the architectural mechanisms, they claim, it is possible extract rich metadata from documents with minimal effort on the part of the end user.

From a single day’s discussion, one could hardly expect a resolution of complex technical issues. However, the participants seemed to agree that useful discussions had begun, and that the approaches presented deserved follow-up efforts. At this point, there is a lively discussion going on between a couple of leaders from the HyTime and RDF camps, but no decision has been taken on whether to hold a second meeting, much less to capture any results in a formal standard.

Conclusion and Recommendations

The world of SGML appears to be quite healthy, whether one looks at the fundamental level of standards development or surface layers of application.

Although DOE has been involved with SGML and its predecessors since the late 1970s, interest in these subjects has tended to reside in specialized groups. The rise of the WWW brought a casual, if frequently effective, use of SGML (in the form of HTML) to a wide community but did not spread wide understanding of the underlying technology. The rise of XML and its adoption by major software houses suggests that use will become even more widespread. For some uses, a casual approach to XML may suffice. However, for records, product data, and other mission-sensitive information, DOE should take an active position on the development and use of SGML-related standards.

Because DOE is one of those organizations adopting WG4 standards, it should continue active participation in WG4’s work, particularly the revision of SGML and the application of DSSSL and

HyTime. As DOE's use of these standards increases, the need for continued commitment to their maintenance and extension will increase as a consequence. DOE should also keep aware of developments in the realm of applications by participating in conferences and developers' groups. Furthermore, DOE should establish more internal means for sharing tools, techniques, and applications. Extension of the NWIG metadata system and construction of a comprehensive records system such as that proposed by the Y-12 WRAP project can profit from future support of SGML/XML by DOE. Y-12, as the leader in development of SGML-related standards, is in a good position to continue also as a leader in their application.

Future meetings

WG4 has the following meetings scheduled for the next year:

Group	Dates	Location	Host
WG4	11–15 November 1998	Chicago	GCA
JTC1	January 1999	Brazil	
WG4	April 1999	Europe	GCA

There may also be project meetings between WG4 meetings.

WG4 has started scheduling most of its meetings in conjunction with conferences sponsored by the GCA. These conferences generally deal with SGML, HyTime, DSSSL, and related topics; combining meetings with them allows a reduction in the number of trips for experts who participate in both activities. My travel to this meeting was supported in part by GCA.